Lecture 36

In this supplementary lecture we discuss further the distinction between population parameters and sample statistics.

If X is a random variable associated with a population, then the population parameters of interest to us are:

- the population size N;
- the population mean $\mu = \mu_X = E(X) = \dfrac{\sum\limits_{x \in X} x}{N}$;
- the population variance $\sigma^2 = \sigma^2{}_X = Var(X) = \dfrac{\sum\limits_{x \in X} (x - \mu)^2}{N}$; and
- the population standard deviation $\sigma = \sigma_X = \sqrt{Var(X)}$.

When the distribution of values of X for a theoretical or an actual population can be represented by a probability function $Pr(X = x)$, we can use the probabilities to calculate μ and σ:

- $\mu_X = E(X) = \sum xPr(X = x)$; and
- $\sigma^2{}_X = E(X^2) - [E(X)]^2$ where $E(X^2) = \sum x^2 Pr(X = x)$.

The situation is a bit different for a sample of X-values $\{x_1, x_2, \ldots, x_n\}$. The sample statistics of interest to us are:

- the sample size n;
- the sample mean $\overline{X} = \dfrac{x_1 + x_2 + \cdots + x_n}{n}$;
- the sample variance $s^2 = s^2{}_X = \dfrac{(x_1 - \overline{X})^2 + (x_2 - \overline{X})^2 + \cdots + (x_n - \overline{X})^2}{n - 1}$; and
- the sample standard deviation $s = s_X = \sqrt{\text{sample variance}}$.

Notice that in the formula for the sample variance the sum of the squared deviations from the mean is divided by $(n - 1)$ rather than n. Statisticians have found, from experience, that the quotient gives a more informative and useful measure of the spread of the data if $(n-1)$ is used as the denominator instead of n.

We give an example below that illustrates the use of these sample statistics.

Suppose that a die is cast. Here is a familiar table, from which the expected value and the variance may be obtained.

X	X^2	$Pr(X = x)$	$xPr(X = x)$	$x^2Pr(X = x)$
1	1	1/6	1/6	1/6
2	4	1/6	2/6	4/6
3	9	1/6	3/6	9/6
4	16	1/6	4/6	16/6
5	25	1/6	5/6	25/6
6	36	1/6	6/6	36/6
		1	21/6	91/6

So $E(X) = 21/6 = 3.5$ and $E(X^2) = 91/6 \approx 15.17$. Then $Var(X) \approx 15.17 - (3.5)^2 = 2.92$. So $\sigma_X \approx \sqrt{2.92} = 1.71$.

Now suppose that a sample of 10 values of X is obtained. In other words, suppose that a die is tossed 10 times. Here is one possible set of outcomes: $\{6, 6, 2, 4, 1, 3, 4, 6, 2, 4\}$

The sample size is 10. We can construct a table to help us calculate the other sample statistics:

X	$\lvert x - \overline{X}\rvert$	$\lvert x - \overline{X}\rvert^2$
6	2.2	4.84
6	2.2	4.84
2	1.8	3.24
4	0.2	0.04
1	2.8	7.84
3	0.8	0.64
4	0.2	0.04
6	2.2	4.84
2	1.8	3.24
4	0.2	0.04
38		29.60

We see that $\overline{X} = 38/10 = 3.8$, and $s^2 = 29.60/9 \approx 3.29$ so that $s \approx 1.81$.

Notice that the sample statistics $\overline{X} = 3.8$ and $s_X = 1.81$ differ somewhat from the population parameters $\mu_X = 3.5$ and $\sigma_X = 1.71$. This is to be expected. If we increase the sample size, we are more likely to get sample statistics that are close to the population parameters. Also, if we average the means from a large number of samples then we can expect that the result will be close to the population mean. The details can be found in a more advanced course on statistics.

This ends the course.